Deep Learning for Multi-View Cancer Drug Response Prediction

Boyi Chen[1], Jianwei Huang[1], Yintao Cai[2], Yuetong Jiang[1], Zelong Yang[1]

[1] Artificial Intelligence Research Institute, Xiamen University
[2] School of Information, Xiamen University
36920231153181,36920231153195,23020231154165,36920231153199,36920231153256

Abstract

Cancer treatment remains a central challenge in the field of medicine. Cancer drug response prediction is a fundamental task at the intersection of medicine and computer science, which opens up opportunities and possibilities that can significantly impact cancer treatment. Currently, various model methods have been developed to select drugs based on cancer and cell line types, enhancing the efficiency of cancer treatment. However, the challenge of missing drug response values due to unknown cancer or tumors remains unsolved, posing a significant hurdle to cancer treatment. In response to this, we propose a model to predict drug response values. This model will initially apply various feature representations to the input data, generating corresponding embeddings. Then, the predicted drug response values are generated according to the views which are created to examine the interrelationships among these representations. Finally, we conduct our experiments on real data sets and give relevant experimental conclusions to verify the performance and accuracy of our model.

Introduction

At present, many methods have been proposed for the treatment of major diseases such as cancer and tumors, offering new opportunities and possibilities in the realm of disease treatment(Verma 2012). Nevertheless, there are still challenges in the treatment process, exerting a significant impact on its efficacy(Zugazagoitia et al. 2016).The variability of cell lines underscores the need for customized drug treatments that may differ across various types of cancer or cell lines (Baptista, Ferreira, and Rocha 2021).The heterogeneity in drug response presents a formidable challenge in the treatment process, as even identical drugs can produce different outcomes when administered to diverse patients or cancer types. It not only fails to yield the desired therapeutic effect but may also worsen the patient's condition. Furthermore, many drugs cannot be universally effective for various types of cancer experienced by patients.

Given the availability of numerous public data sets, particularly clinical data sets, including but not limited to cell line databases, cancer genome databases, cancer drug databases, etc., it provides an opportunity for us to propose or enhance the existing cancer drug prediction models. This innovative approach to cancer treatment holds the potential to significantly advance cancer treatment research, offering greater prospects for the treatment of major diseases such as cancer or tumors.

Currently, many models have been proposed, which use cell line information and drug molecular information as inputs to predict the drug response (CDR) value of cancer (Ye et al. 2021). This also includes machine learning models (Chen and Zhang 2021) such as Ridge Regression (Geeleher, Cox, and Huang 2014), MOLI (Sharifi-Noghabi et al. 2019), and DeepCDR (Liu et al. 2020).

In fact, in the treatment of unknown cancers or tumors, the absence of CDR values is a common occurrence, often attributed to the natural variability of these responses. This phenomenon is also recognized as a manifestation of the challenge known as "missing view values" in multi-view learning (Chao and Sun 2019). However, the lack of CDR values seriously increases the difficulty of the treatment of this cancer or tumor, which means that we need a predictive model with robust performance and high accuracy. It can predict the CDR value of different drugs based on unknown cancer or tumor. The greater the accuracy of the predictive model, the more effective the treatment for the cancer or tumor becomes. However, there is still room for improvement in the performance or efficiency of current methods, because they are still missing or inaccurate in predicting CDR value of certain types of cancer.

In summary, our primary objective is the prediction of Cancer Drug Response (CDR) values. The motivation for selecting this problem stems from the current convergence of cancer drug response prediction and machine learning, which has evolved into a burning issue recently. As we see, many machine learning models have been proposed to predict the drug response value of cancer. From a more macroscopic perspective, research in this particular domain offers some key advantages: it can propel advancements in the medical treatment of cancer, enhance the efficacy of treatment, improve patient-specific drug selection, reduce duration of treatment, among others. These inherent benefits make this a compelling area of study with significant implications.

Related work

The major challenge in drug response prediction lies in the fusion of heterogeneous data from multiple sources. Technically, the significant differences in the data structure, dimensionality, signal-to-noise ratio, and complexity of multi-omics data pose a great challenge in representation learning.

Deep learning methods were generally based on combinations or stacks of Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and attention mechanisms. MVLR (Ammad-Ud-Din et al. 2017) proposed a multi-view multi-task model based on functional link networks, which treated different data sources as independent feature sets. CDRScan (Chang et al. 2018) used stacked CNNs to map inputted cell line data and molecular drug data to their corresponding CDR value. MOLI (Sharifi-Noghabi et al. 2019) used triplet loss to receive the integrated outputs of several different neural networks, where a network was applied to a kind of input data. DeepCDR (Liu et al. 2020) utilized graph convolution network (GCN) to process drug molecular information, and merged it with cell line multi-omics data to predict CDR values. DeepDSC (Li et al. 2019) used stacked encoders to extract genomics from gene expression data, and then chemical characteristics of drugs were jointed to generate response values. (Lee et al. 2022) advanced a "gene-centric multi-channel" (GCMC) model, which employs 3D tensor transformation for a more potent integration of multi-omics data in drug response predictions. Moreover, (Sagingalieva et al. 2023) developed the Hybrid Quantum Neural Network (HQNN), indicating a 15% improvement in predictability for IC50 values over traditional methods, thereby enhancing training efficiency. Meanwhile, (Liu, Tong, and Chen 2023) introduced a multi-view method that efficiently leverages multi-view information through a self-attention mechanism coupled with multi-scale fusion, significantly augmenting feature representations.

A major contribution is the GraphCDR model as proposed by (Liu et al. 2022), which leverages a graph neural network to amalgamate multi-omics profiles of cancer cell lines, drug compositions, and disclosed responses. The model uniquely integrates a contrastive learning task utilized as a regularizer, effectively optimizing the model's learning capability. In scenarios where validated responses are absent, GraphCDR proves its worth by adeptly utilizing biochemical information in accurately predicting CDR for uncharted cell lines and drugs. An additional valuable contribution is the automated Cancer Drug Response Prediction framework via a Graph Neural Network (GNN) model, known as AutoCDRP, as published by (Oloulade et al. 2023). This model employs a surrogate model tailored to estimate the performance of GNN structures randomly selected from a predefined search space, thereby facilitating the selection of the most ideal architecture based on evaluation performance. Furthermore, (Wang et al. 2023) developed the XMR model, an innovative multimodal neural network model, constructed via a fusion of a visible and a graph neural network. This model, specifically conceived for predicting drug responses targeting Triple-Negative Breast Cancer, acknowledges the interconnectedness of genomic attributes and drug structural features. The visible neural network encapsulates the genomic characteristics, whilst the graph neural network administers the drugs' chemical composition.

Collectively, these studies underline the profound potential of neural networks in drug response prediction, providing fresh pathways for refining and customizing cancer treatments. However, these methods commonly ignore the view value missing problem from unknown cell lines of cancers or tumors. We plan to fuse multi-omics data in a multi-view framework to alleviate the view missing problem and generate accurate predictions.

Method

This section introduces BoyNet. BoyNet consists of three stages: input data representation, view generation, and view combination. The general framework of BoyNet is shown in Figure 1.

Firstly, all input data are transformed into latent space using an embedding representation, as these multi-omics data and drug feature data is heterogeneous and cannot be interconnected to connect their corresponding response values. The data representation operation is displayed to the left of Figure 1. Secondly, generate several views to receive embedded features, and then calculate potential interactions. The goal of these views is to observe potential reactions between cell lines and drugs through embedding or feature observation. These views are displayed in the middle of Figure 1.

Finally, combine these views and pass them through linear layers to obtain a CDR prediction map, and compare it with the true values for backpropagation; this operation is displayed on the right side of Figure 1.

Molecular data representation

Mutation embedding Mutation data is a sequence composed of $\{0, 1\}$ symbols, with mutation locations marked by symbol 1. The definition of mutation embedding process is as follows: $\boldsymbol{E^m} = F(\boldsymbol{M})$, where Ne represents the size of embedding, N_x is the maximum number of mutation positions in the cell line, $\boldsymbol{M} \in \mathbb{R}^{N_c \times N_m}$ is a mutation sequence, $\boldsymbol{E^m} \in \mathbb{R}^{N_c \times N_x \times N_e}$ represents mutation embedding, whereN_c represents the number of cell lines, and N_m represents the total number of mutation sites.

Genetic expression features The gene expression characteristics are composed of a series of continuous values. The definition of embedding operation is as follows: $\boldsymbol{E^p} = F(\boldsymbol{P})$, $\boldsymbol{P} \in \mathbb{R}^{N_c \times N_p}$ is a mutated sequence, $\boldsymbol{E^p} \in \mathbb{R}^{N_c \times N_e}$ represents compressed expression features, where N_p represents the number of genes in a single cell line.

Methylation features The methylation feature is also composed of a series of continuous variables. The definition of compression operation is as follows: $\boldsymbol{E^j} = F(\boldsymbol{J})$, $\boldsymbol{J} \in \mathbb{R}^{N_c \times N_j}$ is a mutated sequence, $\boldsymbol{E^j} \in \mathbb{R}^{N_c \times N_e}$ represents the compressed methylation feature, where N_j represents the number of methylation sites.

Cell line latent features and drug latent features Decompose the CDR matrix into potential features of cell lines and potential features of drugs. The operation definition is as follows: $\boldsymbol{E^c}, \boldsymbol{E^d} = F(\boldsymbol{R})$, where R is the observed CDR

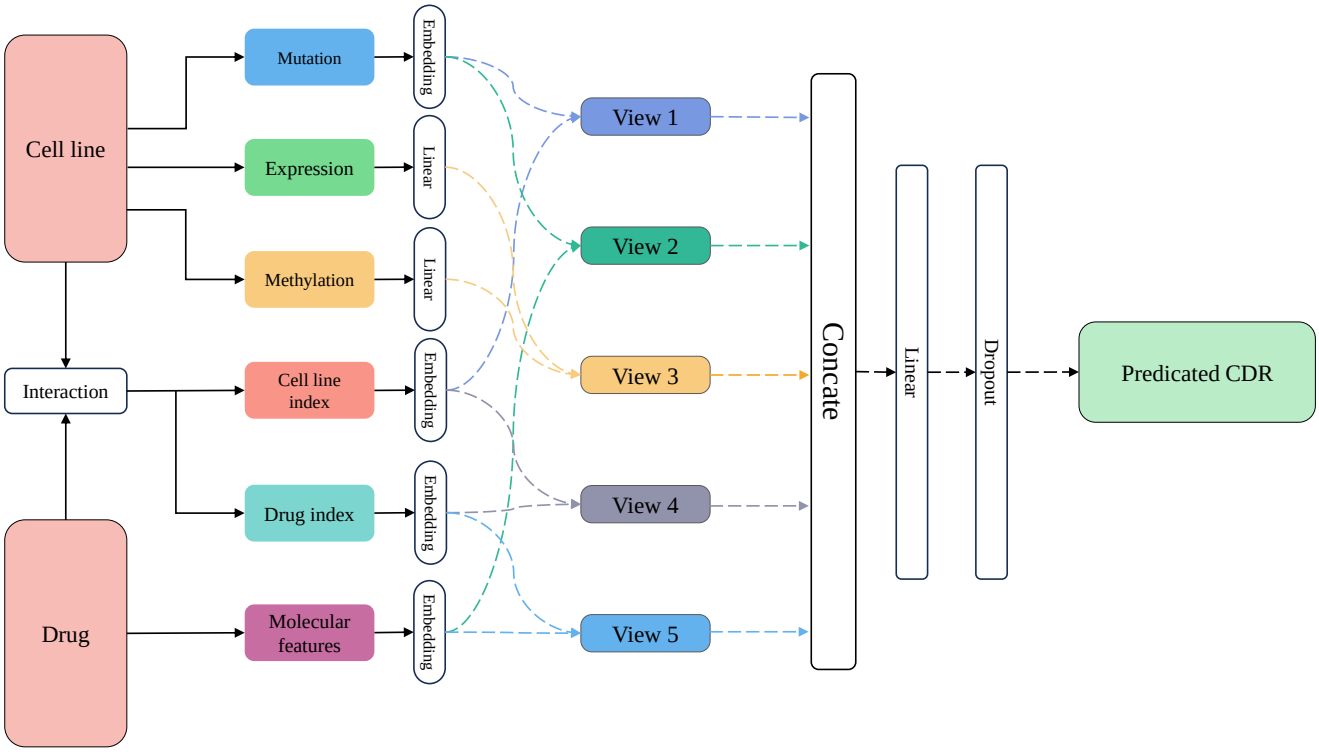


Figure 1: The graphical illustration of the proposed **B**iological Multi-**O**mics S**Y**nergy **Net**work(BoyNet)

value, $\boldsymbol{E^c} \in \mathbb{R}^{N_c \times N_e}$ is a potential feature of the cell line, $\boldsymbol{E^d} \in \mathbb{R}^{N_d \times N_e}$ represents the potential feature of the drug, where Nd represents the number of drugs.

Drug molecular embeddings Drug molecules are composed of multiple atoms, each with similar characteristics, which is also a sequence of $\{0, 1\}$ symbols. Molecules are also represented by embedding and are represented as follows: $\boldsymbol{E^f} = F(\boldsymbol{F})$, where $\boldsymbol{F} \in \mathbb{R}^{N_d \times N_a \times N_f}$ is a molecular characteristic tensor, $\boldsymbol{E^f} \in \mathbb{R}^{N_d \times N_a \times N_f \times N_e}$ is a mutation embedding.

View generation

In order to observe the interaction between cell line information and drug information, five views were generated to learn the potential relationships between multiple sets of data

View 1 Learning the interconnection between cell lines and their mutations. Given a cell line c, the result of view 1 is represented as follows:

$$O^1 \leftarrow V1\left(\boldsymbol{E^m_{c,:,:}}, \boldsymbol{E^c_{c,:}}\right) \tag{1}$$

View 2 Learning the relationship between genomic mutations and drug molecular characteristics. Given a cancer drug pair (c, d), the results of View 2 can be represented as follows:

$$O^2 \leftarrow V2\left(\boldsymbol{E^m_{c,:}}, \boldsymbol{E^f_{d,:,:,:}}\right) \tag{2}$$

View 3 learns the connection between expression and methylation. The result of view 3 can be represented as follows:

$$O^3 \leftarrow V3\left(\boldsymbol{E^p_{c,:}}, \boldsymbol{E^j_{c,:}}\right) \tag{3}$$

View 4 The relationship between the cell line latent matrix and the drug latent matrix for learning decomposition. The result of view 4 can be represented as follows:

$$O^4 = \boldsymbol{E^c_{c,:}} \cdot \boldsymbol{E^d_{d,:}} \tag{4}$$

View 5 learns the interconnection between potential drug features and drug molecular features. The result is represented as:

$$O^5 \leftarrow V5\left(\boldsymbol{E^d_{d,:}}, \boldsymbol{E^f_{d,:,:,:}}\right) \tag{5}$$

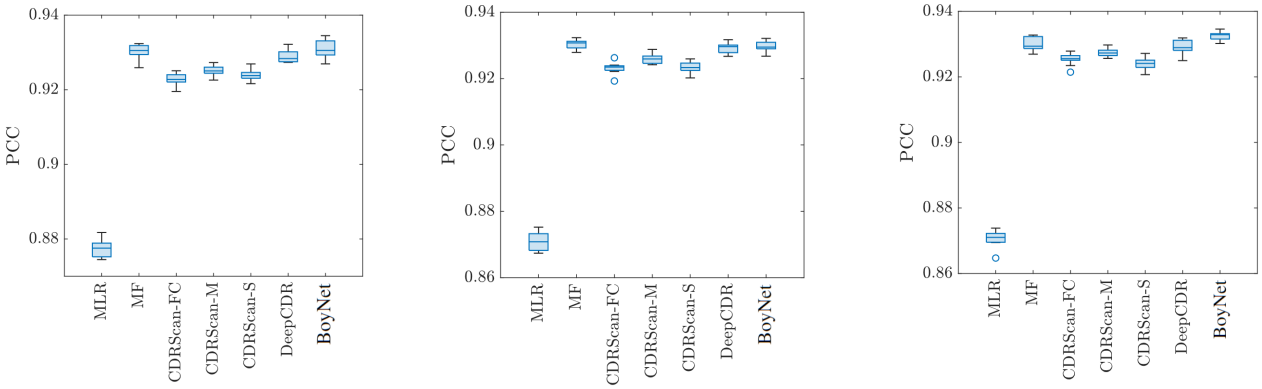
We use inner product to implement all view generation processes

View combination

We collected the outputs of the five views mentioned above and obtained

$$O = \left[O^1, O^2, O^3, O^4, O^5\right] \tag{6}$$

Where $O \in \mathbb{R}^{1 \times (N_x + N_a * N_f * 2 + 2)}$ represents the connected view output, which is finally received by two consecutive linear layers to generate the final prediction.



(a) non-overlapping cell line split. (b) non-overlapping drug split. (c) random split.

Figure 2: Box-plots of seven methods in terms of PCC. (a) split based on non-overlapping cell lines. (b) split based on non-overlapping drugs. (c) random split

Experiment

This section introduces datasets, performance measurements, comparable methods, results and model configurations. The performance competition is provided in the perspective of prediction performance.

datasets

We sourced raw data from eminent bioinformatics databases, including the Cancer Cell Line Encyclopedia (CCLE), which provides a comprehensive catalogue of cancer cell line models. Our analysis was centered on multi-omics data for three specific cell lines, encompassing gene expression, methylation, and mutation. Furthermore, we utilized The Cancer Genome Atlas (TCGA), an extensive repository of human cancer genomes, which includes data on mutations, mRNA, miRNA expression, and methylation. Additionally, the Genomics of Drug Sensitivity in Cancer (GDSC) served as a critical public resource, offering a vast array of IC50 values that correlate with cellular and drug pairings. The chemical database PubChem, which is the largest of its kind globally, supplied structural data for a multitude of drugs. Our dataset comprised data from 494 distinct cell lines, 237 drugs, and 94,314 observed response measures, with approximately 29% of the data missing. We collated comprehensive gene expression data (494 × 697 dimensions), methylation data (494 × 808 dimensions), and gene mutation data (494 × 34673 dimensions).

Competition models

1. Matrix Factorization (MF) (Wang, Chen, and He 2018) a well-known method commonly applied in forecasting user-item ratings within recommender systems , is utilized here for the prediction of CDR values, i.e., $ln(IC_{50})$.

2. Multiple Linear Regression (MLR) (Geeleher, Cox, and Huang 2014) establishes linear relationships by linking each element of the input with the corresponding output elements. These input elements comprise gene expression, methylation, genetic mutations, and characteristics of drugs.

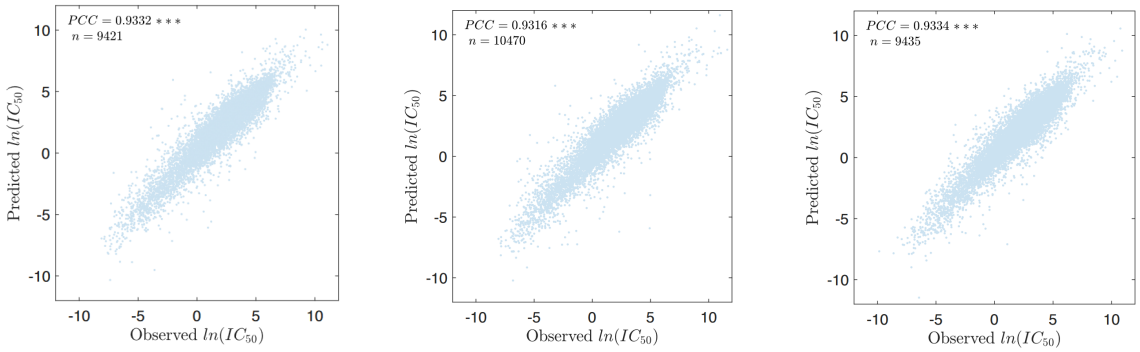
3. CDRScan (Chang et al. 2018) published several versions[1]. The three versions with relative good performance were employed as comparable methods. CDRScan-Master employs a pair of stacked CNNs to process molecular fingerprints and genomic mutations, thereby encoding drug and cancer profiles. These encoded representations are then inputted into a third stacked CNN. In contrast, CDRScan-Shallow features fewer layers in this third CNN stack, opting instead to incorporate a greater number of linear layer operations. Meanwhile, CDRScan-FullConnected substitutes the third stacked CNN entirely with fully connected layers. For convenience, CDRScan-Master, CDRScan-Shallow, and CDRScan-FullConnected are presented by CDRScan-M, CDRScan-S, and CDRScan-FC, respectively.

4. DeepCDR (Liu et al. 2020) leverages CNNs and GCNs to process multi-omics data of cell lines and chemical features of drugs, respectively. The codes are available at github.com[2].

model configurations

For fair competition on all models, the batch size is set to 64. The Adam (Kingma and Ba 2015) optimizer is adopted to train all the models. All the compared models are implemented in PyTorch 1.8.2 (LTS), and are ran four graphics processing units of NVIDIA Tesla V100. Moreover, the comparable models have achieved the same accuracy as their corresponding literature. All the size of latent features and the size of embeddings are set to 20. The MF is solely based on interaction data, i.e., IC50. The MLR, DeepCDR, CDRScan use all the molecular level data as input and cor-

[1]http://github.com/summatic/CDRScan
[2]http://github.com/kimmo1019/DeepCDR



(a) non-overlapping cell line split. (b) non-overlapping drug split. (c) random split.

Figure 3: The visualized correlations between observed CDR values and predicted CDR values in terms of three train/test split methods.

responding observed IC50 as target. All these methods are running on the same 10-fold cross-validation.

comparisons

A variety of metrics are employed to gauge the predictive performance of the $ln(IC_{50})$ values. In this study, we utilized a widely recognized measure: the Pearson Correlation Coefficient (PCC), as suggested by Liu et al.

$$\text{PCC} = \frac{\sum_{(c,d)\in \boldsymbol{T}} \left(R_{c,d} - \bar{R}\right)\left(\hat{R}_{c,d} - \hat{\bar{R}}\right)}{\sqrt{\sum_{(c,d)\in \boldsymbol{T}} \left(R_{c,d} - \bar{R}\right)^2}\sqrt{\sum_{(c,d)\in \boldsymbol{T}} \left(\hat{R}_{c,d} - \hat{\bar{R}}\right)^2}} \tag{7}$$

where $\boldsymbol{T}$ is the testing set, $R_{c,d}$ is a real value, $\hat{R}_{c,d}$ is a predicted value, $\bar{R}$ is the mean value in the testing set, and $\hat{\bar{R}}$ is the mean value of predicted CDR values.

For fair competition, all the comparable methods were ran on common tasks. The experimental results are shown in Figure 2. When observing at the metric on three groups of data, the proposed BoyNet has the best performance for all metrics when compared with other methods. Compared with BoyNet, MF ignores the intrinsic characteristics of drugs and cell lines, the predictions are generated based on the learned latent features. This reveals the impact of molecular data. For CDRScan-FC, more linear layers bring better stability but may affect the model's ability to capture nonlinear features. For CDRScan-S, fewer convolutional units reduce the accuracy. Compared with CDRScan, DeepCDR simplifies the processing of omics data, which can fuse more omics data. DeepCDR achieves third-best performance, somehow owning to the benefits from GCN on drug molecular representation. There is still a gap between DeepCDR and MF. A possible reason for this improvement is that latent interactions are much more important than feature representation only. BoyNet shows the best performance. It uses an embedding component to encode the input data, which compresses the high-dimensional input data into a low-dimensional feature space and filters out fluctuations. This allows BoyNet to quickly and efficiently complete the extraction of low- and high-order latent interactions.

prediction analyses

Three groups of validation experiments on the proposed BoyNet were visualized using scatter plots, see Figure 3. The CDR values of new cell lines are visualized in Figure 3a. The CDR values of new drug are visualized in Figure 3b. The CDR values of known cell lines and known drugs are visualized in Figure 3c. Due to the lengthy nature of displaying the plots for 10-fold predictions, we have selected a random single result from each set of the 10-fold datasets for presentation.

When examining the PCC values across the three subfigures, it is evident that they maintain a narrow range of variance. The devised BoyNet technique demonstrates a proficient capability in addressing the challenge of missing view values. The substantial correlation exhibited by the PCC values between the actual and forecasted values signifies the BoyNet method's superior predictive precision. One plausible explanation for this phenomenon could be that elevated concentrations are indicative of diminished performance. Across all examined datasets, responses that are classified as ineffective outnumber those that are effective.

Conclusions

Our work on the Biological Multi-Omics Synergy Network (BoyNet) represents a step forward in the field of cancer-drug-response (CDR) prediction. BoyNet introduces a method for addressing the challenge of missing CDR values in less studied cancer types or tumors by employing equal-dimensionality embeddings to create a more comprehensive data analysis framework. Our results show promise in enhancing the accuracy of drug response predictions for a range of cancer cases, which may assist in the development of more tailored treatment plans for patients. BoyNet will serve as a useful reference for further research targeting the optimization of tumor treatment approaches.

References

Ammad-Ud-Din, M.; Khan, S. A.; Wennerberg, K.; and Aittokallio, T. 2017. Systematic identification of feature combinations for predicting drug response with Bayesian multi-view multi-task linear regression. *Bioinformatics*, 33(14): i359–i368.

Baptista, D.; Ferreira, P. G.; and Rocha, M. 2021. Deep learning for drug response prediction in cancer. *Briefings in bioinformatics*, 22(1): 360–379.

Chang, Y.; Park, H.; Yang, H.-J.; Lee, S.; Lee, K.-Y.; Kim, T. S.; Jung, J.; and Shin, J.-M. 2018. Cancer drug response profile scan (CDRscan): a deep learning model that predicts drug effectiveness from cancer genomic signature. *Scientific reports*, 8(1): 8857.

Chao, G.; and Sun, S. 2019. Semi-supervised multi-view maximum entropy discrimination with expectation Laplacian regularization. *Information Fusion*, 45: 296–306.

Chen, J.; and Zhang, L. 2021. A survey and systematic assessment of computational methods for drug response prediction. *Briefings in bioinformatics*, 22(1): 232–246.

Geeleher, P.; Cox, N. J.; and Huang, R. S. 2014. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome biology*, 15: 1–12.

Kingma, D.; and Ba, J. 2015. Adam: A method for stochastic optimization in: Proceedings of the 3rd international conference for learning representations (iclr'15). *San Diego*, 500.

Lee, M.; Kim, P.-J.; Joe, H.; and Kim, H.-G. 2022. Gene-centric multi-omics integration with convolutional encoders for cancer drug response prediction. *Computers in Biology and Medicine*, 151: 106192.

Li, M.; Wang, Y.; Zheng, R.; Shi, X.; Li, Y.; Wu, F.-X.; and Wang, J. 2019. DeepDSC: a deep learning method to predict drug sensitivity of cancer cell lines. *IEEE/ACM transactions on computational biology and bioinformatics*, 18(2): 575–582.

Liu, Q.; Hu, Z.; Jiang, R.; and Zhou, M. 2020. DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. *Bioinformatics*, 36(Supplement_2): i911–i918.

Liu, X.; Song, C.; Huang, F.; Fu, H.; Xiao, W.; and Zhang, W. 2022. GraphCDR: a graph neural network method with contrastive learning for cancer drug response prediction. *Briefings in Bioinformatics*, 23(1): bbab457.

Liu, Y.; Tong, S.; and Chen, Y. 2023. HMM-GDAN: Hybrid multi-view and multi-scale graph duplex-attention networks for drug response prediction in cancer. *Neural Networks*, 167: 213–222.

Oloulade, B. M.; Gao, J.; Chen, J.; Al-Sabri, R.; and Wu, Z. 2023. Cancer drug response prediction with surrogate modeling-based graph neural architecture search. *Bioinformatics*, 39(8): btad478.

Sagingalieva, A.; Kordzanganeh, M.; Kenbayev, N.; Kosichkina, D.; Tomashuk, T.; and Melnikov, A. 2023. Hybrid quantum neural network for drug response prediction. *Cancers*, 15(10): 2705.

Sharifi-Noghabi, H.; Zolotareva, O.; Collins, C. C.; and Ester, M. 2019. MOLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics*, 35(14): i501–i509.

Verma, M. 2012. Personalized medicine and cancer. *Journal of personalized medicine*, 2(1): 1–14.

Wang, Z.; Chen, K.; and He, L. 2018. Asysim: Modeling asymmetric social influence for rating prediction. *Data Science and Pattern Recognition*, 2(1): 25–40.

Wang, Z.; Zhou, Y.; Zhang, Y.; Mo, Y. K.; and Wang, Y. 2023. XMR: an explainable multimodal neural network for drug response prediction. *Frontiers in Bioinformatics*, 3.

Ye, Q.; Hsieh, C.-Y.; Yang, Z.; Kang, Y.; Chen, J.; Cao, D.; He, S.; and Hou, T. 2021. A unified drug–target interaction prediction framework based on knowledge graph and recommendation system. *Nature communications*, 12(1): 6775.

Zugazagoitia, J.; Guedes, C.; Ponce, S.; Ferrer, I.; Molina-Pinelo, S.; and Paz-Ares, L. 2016. Current challenges in cancer treatment. *Clinical therapeutics*, 38(7): 1551–1566.